**February 29, 2012**

**Tuuli Lappalainen, tuuli.lappalainen@unige.ch**

**University of Geneva**

**Geuvadis RNA sequencing guidelines**

Updated version of the Geuvadis RNAseq sequencing protocols

**Table 1: Sample allocation to Geuvadis partners**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Lab #** | **mRNAseq lib prep** | **mRNAseq sequencing** | **miRNAseq lib prep & seq** |
| UNIGE (Geneva) | 1 | 465 | 116 full coverage,  168 half coverage | 116 |
| CRG (Barcelona) | 2 | 96 | 96 | 96 |
| MPIMG (Berlin) | 3 | 72 | 72 | 72 |
| ICMB (Kiel) | 4 | 72 | 72 | 72 |
| HMGU (Munich) | 5 | 48 | 48 | 48 |
| UU (Uppsala) | 6 | 48 | 48 | 48 |
| LUMC (Leiden) | 7 | 48 | 48 | 48 |

Counts include 5 duplicate samples that will be sequenced in full coverage every lab

**Table 2: RNA samples by population**

|  |  |  |
| --- | --- | --- |
|  | **Geuvadis** | **Geuvadis + 1KG Phase1** |
| TSI | 93 | 92 |
| GBR | 96 | 86 |
| FIN | 95 | 89 |
| CEU | 92 | 79 |
| YRI | 89 | 77 |
| TOTAL | 465 | 423 |

**RNA samples**

- Total RNA was extracted with Trizol in Geneva in October

- Qbit and total RNA Bioanalyzer has been done in Geneva to estimate RNA quality and quantity. The Bionanalyzer result files can be found in: WP4 Wiki > Analysis > QC

- The laboratories have received a minimum of 4 ug of total RNA of the samples allocated to them

- The RNA sample doesn't require further purification steps (other than what is specified in the mRNAseq protocol to get poly-A mRNA)

- All laboratories should do total RNA Bioanalyzer for 10-20% of the RNA samples before library prep

- Library preps should be done in a random order in every lab. The sample information sheets are already randomized and you can process them in that order – do not sort them according to any parameter.

- The labs may choose freely the input quantity of total RNA to use in the library preps according to what works best in their lab. You should use a small proportion for mRNAseq (we use 0.5 ug) and another part for miRNAseq (we use 1 ug). Keep the 4 ug limit in mind, but it should be enough for some reruns too.

**mRNAseq:**

- All the partners do the library prep of the samples that were allocated to them and sequence their libraries from that prep. No libraries will be shared.

- In addition to this, Geneva does library prep of all the samples that are sequenced in other labs. They sequence 168 of these samples (almost all CEU + YRI that were sequenced in other labs and ~12 form the other populations) in in about 1/2 of the normal coverage.

- The sample allocations include 5 samples that will be done in full coverage replicate in every lab. Geneva preps and sequences these samples twice to have within-lab replicates.

- Sequencing is done with HiSeq

- 75 bp paired-end sequencing with fragment size of 280 bp. If you cannot do 75bp, you can sequence 100bp PE and trim the reads.

- Coverage: minimum 10M mapped and properly paired read pairs. This number should be estimated from reads trimmed to 50bp and paired-end mapped with bwa (or an equivalent mapper) to the reference genome (hg19), without filtering for mapping quality.

- Library prep kit: TruSeq RNA Sample Prep Kit v2, the high-throughput protocol

- Cluster generation kit: TruSeq PE Cluster Kit v3

- Sequencing kit: TruSeq SBS Kit v3

- The partners can choose freely how to pool the samples to get the desired coverage

- Bioanalyzer should be run for all library preps before sequencing

**miRNAseq**

- All the partners will do the library prep and sequence that in full coverage (3 M total reads) of the samples that were allocated to them

- These allocations include 5 samples that will be done in full coverage every lab. Geneva preps and sequences these 5 samples only once (no within-lab replication).

- Sequencing is done with HiSeq

- 36 bp single-end sequencing (or trimmed to that). Fragment size 145-160 bp.

- The initial goal was to get 3M total reads per sample from a good quality experiment. The miRNA yield turned out to be worse than expected, and estimating what a good quality experiment actually is not straightforward, so the criteria were refined: the labs should aim either for >3M reads per sample, or a median >6M. As a simple quality assessment, each sample should have around 100 000 reads mapping to human mature miRNA sequences after adaptor trimming, but this is not a shard cutoff but a general guideline.

- In UNIGE we are doing the ethanol precipitation step after gel elution, but if the protocol works well in your lab without it, we don’t require it.

- Library prep kit: TruSeq Sm RNA Sample Prep

- Cluster generation kit: TruSeq SR Cluster Kit v3

- Sequencing kit: TruSeq SBS Kit v3

- The attached bioanalyzer file shows an example of successful library prep results from Geneva - the samples marked with yellow are from Geuvadis and the blue ones are other LCL samples.

**Sample processing information**

- All the labs need to fill in information of sample processing for both mRNA and miRNAseq in order to enable control of batch effects. These information is given with the sample information sheets, which have been emailed to everyone and can also be found under the Samples section of the wiki.

**Raw data processing**

- Each lab needs to submit one demultiplexed fastq file per sample per mRNAper and miRNAseq, produced by CASAVA 1.8 (1.8.2 is OK) allowing one mismatch in the index

- The fastq files for should be named as: SAMPLE\_ID.SeqLabNumber.M/MI\_YYMMDD\_Lane\_Read.fastq.gz . SeqLabNumber has been given in the sample information sheets. M/MI stands for mRNA or miRNA sequencing. YYMMDD is the sequencing date. Example: NA12345.1.MI\_111215\_3\_1.fastq.gz would be sample NA12345 miRNA-sequenced in Geneva on 2011-Dec-15 on lane 3.

- If you've sequenced the same library on several lanes, use the first of them in the filename, but include information of both (e.g. "1\_2") on the sample information sheet. The same applies if the same libraries have been sequenced in several runs (e.g. in miRNAseq to get additional coverage) - but merge these files only after verifying that e.g the quantification results are consistent.

**Data submission**

- The fastq files are to be uploaded to our ftp site (Natalja will submit them later to EBI ENA). Instructions for this can be found form the data storage page of the wiki

**Data submission deadline**

- The FINAL fastq submission deadline is March 15, 2011

- Please submit data when it's done - don't wait for the deadline. We don't want to transfer and map 2 x 500 files at the same time.

- If there are unexpected delays in your data production, inform the other partners (especially Geneva) as soon as possible